Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 82 (2016) 65 – 71

Procedia Computer Science

Symposium on Data Mining Applications, SDMA2016, 30 March 2016, Riyadh, Saudi Arabia

Predicting Critical Courses Affecting Students Performance: A Case Study

Yasmeen Altujjar, Wejdan Altamimi, Isra Al-Turaiki*, Muna Al-Razgan

Information Technology Department
College of Computer and Information Sciences
King Saud University
Riyadh 12372, Saudi Arabia

Abstract

Predicting student academic performance is one of the important applications of educational data mining. It allows academic institutions to provide appropriate support for students facing difficulties. Classification is a data mining technique that can be used to build prediction models. In this paper, we use the ID3 decision tree induction algorithm to build prediction models for academic performance. Our models are built based on records for female students in the Bachelors program at the Information Technology (IT) department, King Saud University, Riyadh, Saudi Arabia. The results indicate that reliable predictions can be achieved based on the performance of students in second year courses. We also identify key courses that can be used as performance predictors. We believe our findings are useful for decision makers at the IT department.

Peer-review under responsibility of the Organizing Committee of SDMA2016

Keywords: Academic performance; decision tree; ID3; data mining

1. Introduction

Data mining is defined as the process of extracting useful and novel information from large amounts of data[1]. It has been applied to provide useful solutions in many areas such as: business, finance, medicine, and healthcare. A relatively new field of data mining applications is *Educational Data Mining* (EDM). Emerging in 2005, EDM is concerned with developing data mining techniques to discover knowledge from data obtained from educational settings[2]. The main goal of this new field of research is to support decision making in academic institutions by analyzing educational data[3]. The information produced by EDM can be useful to several stakeholders in education. For example, it can help instructors evaluate course structure and teaching strategies. In addition, students can get course recommendations based on their progress. Student advisors can benefit from EDM to predict student performance. Predicting low performing students at early stages allows providing additional support for them. According to Baker

* Corresponding author. Tel.: +0-966-11805-2909.
E-mail address: flower.132@hotmail.com, aspirant-1-@hotmail.com, ialturaiki@ksu.edu.sa, malrazgan@ksu.edu.sa

Peer-review under responsibility of the Organizing Committee of SDMA2016
doi:10.1016/j.procs.2016.04.010

and Siemens[4], EDM methods can be classified into: prediction, structure discovery, relationship mining, distillation of data, and discovery with models.

The applications of data mining in the educational context have witnessed rapid growth. There are many factors that led to the growing interest in educational data mining. With the advances in technology, universities are able to accumulate huge amount of academic and non-academic data about their students. Educational institutions are utilizing many resources, such as Learning Management Systems and Student Information Systems, that are generating volumes of data[5]. Well-developed data mining techniques can play an important role in analysing this data.

In this paper, we use data mining techniques to build a model to predict academic performance. The goal of the study is to reveal the courses affecting low academic performance at the Information Technology department (IT) for female students at the College of Computer and Information Sciences, King Saud University, Riyadh. The model will allow the IT department to make the right decisions to monitor and support students and to enhance the quality of the program.

The rest of the paper is organised as follows. In section 2, we review related work in the prediction of academic performance using data mining techniques. Section 3 describes the dataset used, the data mining technique applied, and the experimental results. Finally, section 4 concludes the paper with our findings and recommendations.

2. Literature Review

Predicting student performance is an important application of educational data mining. There are many contributions in this field using different data mining tools and techniques.

Bhardwaj et al.[6] constructed a model to predict academic performance of students enrolled in the Bachelors of Computer Applications programs in five colleges of Dr. R. M. L. Awadh University in India. The data consisted of 300 records for students including: 226 males and 74 females and featured many factors such as: academic, social, demographic, and psychological attributes. The Naive Bayes classification algorithm was used to build the model. It was found that among the most affecting variable in student academic performance are: student grade in the senior secondary school, place of residence (e.g. town, village), and the language of instruction.

In New Zealand, Carnegie et al.[7] used students results at secondary school to predict their performance during the first year in college. The study focused on predicting success in engineering programs. Several models using J48 decision tree classification algorithm were built for this purpose. The models demonstrated that the *Guaranteed Entry Score* is an informative predictor of performance, but alone is not good enough for prediction. For students in the Engineering and Computer Systems specialisation, the study concluded the importance of calculus and physics at the merit and excellence level.

de Morais et al.[8] used data mining techniques to analyse the performance of students at an English e-learning course. The data was obtained for 120 students, 6592 steps, and 11394 transactions. *K-means* algorithm was applied to cluster students based on their answers. Grouping similar students help in selecting future learning activities for each group based on their performance. Regression analysis was then applied to predict the behaviour of students in each cluster. The study showed that the two variables: number of correct answers and the number of correct first attempts, are important for three out of the five obtained clusters.

Lopez et al.[9] developed two classification models to predict the loss of academic status. Decision tree and Naive Bayes algorithm were used to build the models. The study was based on student data from two undergraduate engineering programs at the Universidad Nacional de Colombia. The data of 1532 students consisted of admission and academic data. The former was obtained from the Admission unit and included: initial academic information, demographic and socio-economic, and academic potential. The academic data was obtained from the Academic Information System and included: student records, academic period, program, and GPA. The first model used admission data to predict the loss of academic status at a particular academic period. The second combined both admission and academic data. The effect of unbalanced classes in the data was considered and recovered by applying a cost-sensitive technique in the decision tree model. Models accuracy was evaluated using different data settings. The results showed that Naive Bayes performed better in terms of inbalanced accuracy (up to 85%) in the fourth period. However, decision trees results were more consistent among all periods, which means it is more reliable.

Hashim et al.[10] applied the C4.5 decision tree classification algorithm to predict the performance of student at Alneelain University in Sudan. Academic data of 124 graduate students was obtained from at the Mathematical

Sciences and Statistics department. The data featured student performance at each year in college. The resulting classifier showed the relationship of student performance in each single year and the performance in the last year. The C4.5 model achieved high accuracy on testing data.

At King Saud University in Saudi Arabia, Al-Saleem et al.[11] used ID3 and J48 decision tree algorithms to predict student grades in specific courses. They obtained 112 records for graduates of the Computer Science Masters program. The evaluation of the two models showed that J48 performed better in term of accuracy as compared to ID3. The authors developed an online system based on J48. The system allows students to predict their performance in future courses. In addition, it helps them to select suitable elective courses.

As discussed above, classification and clustering are among the widely used data mining techniques in student performance prediction. These techniques were important at the early days of educational data mining and they are still widely applicable[2]. Many factors are considered when building prediction models for student performance including: academic, social, demographic, and psychological factors.

In this paper, we focus on predicting student performance at the Information Technology department of female students at King Saud University. The Bachelors of Science in Information Technology has a four-year study plan, with two semesters. The first year is the preparatory year, where students have to take general course in mathematics, English, religion, and communication skills. Starting from the second year (third semester), students start taking computer specialised courses along with general education courses. In order to graduate, students are required to take sixteen mandatory specialised courses and seven elective specialised courses of their choices. Classification will be used in this case study to build a model that can predict low achieving students at early stages. The goal is to provide support for those students. In addition, we would like to identify important courses in the program that are good indicators of student level of achievement.

3. Methodology

3.1. Data Description

The dataset used in this study was obtained from the IT department at King Saud University. The Bachelor program includes three tracks: Data Management, Web Technology and Multimedia, and Networks and Security. Each track has core and elective courses. We obtained the records of students who graduated in the academic year 2013-2014. Each record has the following six information:

- Student ID
- Graduation GPA (out of 5)
- High school score (as percentage)
- General Aptitude Test (GAT) score: measures student's analytical and deductive skills.
- Educational Attainment Test (EAT) score: measures the knowledge that students gained in school.
- Courses: This information shows the courses taken by each student.
  A sample of the original dataset is shown in Figure 1.

3.2. Data Pre-processing

The original dataset we obtained in an un-normalized database table where each row represented a student record. It was necessary to pre-process the data in order to prepare it for analysis. Pre-processing was conducted as follows:

- The collected dataset contained multi-value columns (courses columns). Column values were split using Microsoft Excel with *(/)* as a delimiter. This resulted in each course column being separated into four columns: course name, course grade, grade points, and semester.

	المقررات			درجة التحصيلي	درجة القدرات	نسبة الثانوية العامة	معدل التخرج التراكمي النهائي	الرقم الجامعي
عرب/ب+/	221/101 1/هـ/فز	221/121 5/أ+/نجم	221/206 3/ج/عل			92.52	2.66	422
نبت/أ/	231/101 4.75/أ/نجم	231/103 2/د/ريض	231/201 4.75/أ/قصم			88.48	2.16	423
ريض/د-	242/103 5/أ+/عرب	242/121 5/أ+/نجم	242/206 5/أ+/عل	0	0.01	89.72	3.06	424
احص/	261/131 4.5/ب+/نجم	261/132 4/ب/نجم	261/133 3/ج/نجم	0	0.01	96.68	3.09	426
ريض/	261/103 4/ب/عرب	261/121 4.75/أ/نجم	261/206 3.5/ج+/عل	0	0.01	93.19	2.4	426
عل/أ/5	262/101 4.5/ب+/ريض	262/101 4.5/ب+/سلم	262/102 4.5/ب+/نجم	0	0.01	97.32	2.77	426
عل/أ/5	262/101 5/أ+/ريض	262/101 5/أ+/سلم	262/102 4.5/ب+/نجم	0	0.01	95.93	2.8	426
ريض/ج+/	271/103 3.5/ج+/عرب	271/121 4.5/ب+/نجم	271/206 4.5/ب+/عل			93.68	3.78	427
سلم/ب+/	271/102 4/ب/نجم	271/107 4.5/ب+/تال	271/112 3/ج/عل	0	0.01	97.32	2.79	427
سلم/ب+/	271/102 4.5/ب+/نجم	271/107 4.5/ب+/تال	271/112 4/ب/عل	0	0.01	96.52	3.92	427
سلم/أ/	272/102 2.5/د+/نجم	272/107 4/ب/تال	272/112 1/هـ/عل	0	0.01	94.32	3.52	427
سلم/	272/102 3/ج/نجم	272/107 3.5/ج+/تال	272/112 1/هـ/عل	0	0.01	93.06	2.86	427
ريض/ه	281/103 3/ج/عرب	281/121 4/ب/نجم	281/206 3/ج/عل	0	0.01	93.35	3.16	428
ريض/د+/	281/103 3.5/ج+/عرب	281/121 4.75/أ/نجم	281/206 2.5/د+/عل	0	0.01	96	3.55	428
سلم/	282/102 3/ج/نجم	282/107 3.5/ج+/تال	282/112 1/هـ/عل	0	0.01	93.26	3.08	428
/سا	281/102 ع/نجم/	281/107 ع/تال/	281/112 ع/عل/	0	0.01	98.39	3.55	428
سلم/أ/	281/101 5/أ+/قصد	281/101 4.5/ب+/كمي	281/101 3.5/ج+/نظم			98.16	4.77	428

Fig. 1. Sample of the original dataset.

244 ريض	324 احص	222 تم	221 تم	113 عال	104 فيز	151 ريض	106 ريض	111 عال	150 نجم	150 ريض	140 نهج	140 صحة	140 نجم	140 علم	140 ريض	140 تقن	معدل التخرج التراكمي النهائي	الرقم الجامعي
D+	A	B	B	D+	B+	B	B+	B	A+	A+	A+	A+	B+	A+	A+	A+	Good	429
B	C+	C	C	B	D	D	D	F	B	A+	A+	B+	B+	A	A+	A+	Weak	429
A	A+	A	A	A	B+	B+	C+	A	A+	A+	A+	A+	B+	A	A	A+	Good	429
B+	C	C	B	C	D+	D	B+	F	B	A	A	A+	C+	B+	B+	A+	Weak	429
D+	C+	D+	D	D+	D	F	F	F	A+	B+	B	A	B+	B	B+	B+	Weak	429
C	A+	B	B	A	D+	D+	C+	F	B+	A+	A+	A+	C	A+	A+	A+	Weak	429
A+	A+	B+	A	A	A	A+	A	A	A+	A+	A+	A	A	A	A+	A+	Good	429
D+	C	D	B	F	C+	D	C	D+	B+	A	A+	A	B+	A	A+	A+	Weak	429
C	A+	B	C+	B	C	B	C+	C+	A	A+	A+	A	B+	A	B+	A+	Good	429
C+	B+	C	C+	D+	C+	C+	D	C	B	B	A	B	B	B	A+	A+	Weak	429
F	B	C+	C	D	C+	C	C	C+	A	B+	A	A	B+	A+	B+	A+	Weak	429
B	A+	A+	A	A+	A	C	B+	A	A+	A	A+	B+	A+	A	A	A+	Good	429
F	C+	D	D	F	B	D	D	D+	B	B+	A	A	B	A	A	A+	Weak	429
B+	B	B	B	D+	C	D	B	C	A+	A	A	B	B	B+	A+	A+	Good	429
Null	B+	B	B+	A+	C	C	F	A+	B+	A+	A	B+	A	A	A	A+	Good	429
F	D	D	F	F	D+	C	D	D	A+	B+	A+	B+	B+	A	A+	A+	Weak	429
A+	A+	B+	A	A	B+	A	B	B+	A	A+	A+	A+	B+	A	A+	A+	Good	429
B+	A+	B	B+	C+	B	A	C	B+	A+	A+	A+	A+	A+	A	A+	A+	Good	429
B	A	A	B+	B+	C+	B+	B+	C	A+	B+	A	B+	A	A+	B+	A+	Good	429
D	B+	C+	D+	D	A	B	D	D+	B+	B+	A	B+	B+	B+	A	A+	Weak	429
D+	B+	B	C	C+	A	C	D	B+	A+	A	A	B	B+	B+	A+	A+	Weak	429
B	D+	D	C+	F	C+	F	D	D	A+	B+	A	B+	B+	B+	A	A+	Weak	429
A+	A+	A+	A+	A+	A+	A	C	A+	A+	A+	A+	A+	A	A+	A+	A+	Good	429
D+	C	D	D+	F	F	F	F	F	B+	B	A	B+	B+	B+	B+	A+	Weak	429
B+	C	C	B	C	C+	C	D+	C+	C+	A	A+	A	C+	A	B+	F	Weak	429
D+	B	C	D+	D	B	D+	D+	F	C+	C+	B+	B+	B	C	A	B+	Weak	429

Fig. 2. Sample of the dataset after pre-processing.

- After splitting the courses columns, the semester and grade points columns were removed. We assume that the semester in which the course was offered is irrelevant to student performance. The grade points column was discarded because the grade column provided similar information.
- Then, each course of the IT program was represented in a separate column. The values in each column corresponded to the grades achieved in that course.
- The values of track elective courses were set to Null, if the student did not take these courses.
- If a student failed a course, the fail grade was considered instead of the obtained grade after taking the course again. This is will help us in identifying the critical course that affect students performance.
- The GPA column is considered as the class label. Students with a GPA between 4 and 5, corresponding to grades A+, A, B+, or B, were labeled as Good achievers. Students with a GPA less than 4, corresponding to grades C+,C, D+, D, and F, were labeled as Weak achievers. A sample of the dataset after pre-processing is shown in Figure 2.
- The integer values of GAT, EAT, and high school score were partitioned based on equal-depth partitioning method[1].

IF IT325=$A+$ or A or $B+$ **THEN** GPA=Good

IF IT325=C or $D+$ or D **THEN** GPA=Weak

IF IT325=$C+$ **AND** $EAT > 85$ **THEN** GPA=Weak

IF IT325=$C+$ **AND** $EAT < 85$ **THEN** GPA=Good

Fig. 3. The rules obtained from the decision tree based for all students data.

IF IT221=$A+$ or A or $B+$ **THEN** GPA=Good

IF IT221=$C+$ **AND** CSC113=F **THEN** GPA=Weak

IF IT221=$C+$ **AND** CSC113=C **AND** CSC111=$C+$ **THEN** GPA=Weak

IF IT221=F or D or $D+$**THEN** GPA=Weak

Fig. 4. Rules derived from the decision tree based on performance in the first year of study plan

.

3.3. Data Mining

Classification is a widely used technique for predicting student performance. The basic idea is to group students based on known class labels. The *Iterative Dichotomiser* (ID3)[12] is a well-known tree induction algorithm. It is a greedy algorithm in which a decision tree is constructed in top-down recursive approach. The training dataset is iteratively split into smaller partitions. In each iteration, the ID3 algorithm decides which attribute best splits the dataset. The splitting attribute selection is based on maximising information gain. The ID3 algorithm has the advantage of being simple. The resulting classification model can be easily converted to understandable rules[13].

3.4. Experimental Results

Here, we build several models based on ID3 decision tree algorithm. The models are obtained using RapidMiner, the open source predictive analytics platform[14]. The dataset consists of 100 student records. We used 75 records to train the models and the remaining (25 records) were used for testing.

First, we use all data under examination to build a model. The result indicates that the student grade in IT 325 (Operating systems) along with the EAT score are predictors of academic performance, as shows in Figure 3. The IT 325 is one of the difficult courses in the IT plan. An instructor who taught IT 325 for many times conformed this observation.

To better understand the dataset and enable predictions at different stages we divide the dataset into three groups. Each group consists of the courses offered at the same year of the academic plan. A separate model is then built for each group.

We build a decision tree to predict student performance based on courses offered during the first year of the IT plan. Due to space limitation, in Figure 4 we only show some of the important rules obtained from the decision tree model. The result shows that grade obtained in IT 221 (Computer Organisation and Assembly Language) is an indicator of the student final GPA. All students achieving a grade above B+ are classified as Good achievers. Furthermore, for students achieving a grade of C+ in IT 221, their grade in programming courses (CSC111 and CSC113) can be used to predict their performance.

The model obtained using the performance in second year courses indicates that student grades in IT 211 (Human Computer Interaction and Visual Basic) can be used as an indicator of the student final GPA. Figure 5 shows some of the important rules. All students with grades above A were classified as Good achievers. For the rest of the students, the grade obtained in IT 211 needs to be combined with the grades of IT 224 (Network 1) or IT 324 (Information Security) for better prediction.

IF IT211=$A+$ or A **THEN** GPA=Good

IF IT211=B **AND** IT224=$B+$ **THEN** GPA=Good

IF IT211=$B+$ **AND** IT324=$B+$ **THEN** GPA=Good

IF IT211=$D+$ **THEN** GPA=Weak

Fig. 5. Rules derived from the decision tree based on performance in the second year of study plan.

IF IT434=$A+$ or A **THEN** GPA=Good

IF IT434=$C+$ or C **THEN** GPA=Weak

IF IT434=$B+$ **AND** IT361=B **THEN** GPA=Good

IF IT434=$null$ **AND** IT332=$C+$ **THEN** GPA=Weak

Fig. 6. Rules derived from the decision tree based on performance in the third year of study plan

.

Finally, a model is built based on third year courses. The rules shown in Figure 6 indicate that students achieving a grade of A or A+ in the IT 434 (Data Mining) course are classified as Good achievers. However, for students who did not take the course, their performance can be predicted based on other course including: IT 332 (Distributed Systems), IT 422 (Artificial Intelligence), IT 443 (Advanced Human Computer Interaction), and IT 424 (Networks 2).

3.5. Evaluation

In order to evaluate the obtained models, three performance measures are used: *accuracy*, *precision*, and *recall*. Accuracy refers to the percentage of correctly classified records in the testing dataset. Precision measures the percentage of records that the model classified as Good that are actually Good. Recall measures the true positives recognition rate. These measures are calculated as follows:

$$Accuracy = \frac{TP + TN}{P + N} \tag{1}$$

$$Precision = \frac{TP}{P} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

where, *P* is the number of records labelled as Good. *N* is the number of records labelled as Weak. *TP* is the number of records that were correctly classified as Good. *TN* is the number of records that were correctly classified as Weak. *FN* is the number of records that were misclassified as Weak. Table 1 summarises the evaluation measures for each model. The results shows that the model obtained based on the second year courses is the most accurate model.

4. Conclusions

In this paper, we applied the ID3 classification algorithm to the records of 100 graduates from the IT department Bachelors program. The goal was to predict the performance of students and to identify critical courses in the Bachelor IT program. We developed classification models for each year of the program. Our results suggested that the classification model based on performance in the second year is the most accurate. The student performance in IT 221 and the two programming courses, CSC111 and CSC113, is a great indicator of student level of achievement.

Table 1. Performance evaluation of the obtained models.

Model (year)	Accuracy	Class	Precision	Recall
First	68%	Good	76.47%	76.47%
		Weak	50%	50%
Second	80%	Good	87.50%	82.35%
		Weak	66.67%	75%
Third	76%	Good	82.35%	82.35%
		Weak	62.50%	62.50%

From all the models, we can conclude that the courses which can serve as indicators of student performance are as follows: IT 211 (Human Computer Interaction), IT 324 (Information security), IT 224 (Networks 1), IT 434 (Data mining), IT 422 (Intelligent Systems), IT 443 (Advanced Human Computer Interaction), IT 325 (Operating Systems), and IT 424 (Networks 2).

Based on the observations in this study many actions can be taken by the IT department to help students, academic advisers, and instructors. Predicting performance at early stages enables the allocation of proper support to students and eliminate problems close to graduation. The department can focus on the identified courses to understand the difficulties faced by students. The most challenging part of this study was the data pre-processing step. Thus, only 100 student records were used to construct the classification models. In the future, we plan to obtain more student records and to better automate the data pre-processing task.

References

1. Han, J., Kamber, M.. *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann; 3rd ed.; 2011. ISBN 978-0-12-381479-1.
2. Baker, R.S., Inventado, P.S.. Educational Data Mining and Learning Analytics. In: Larusson, J.A., White, B., editors. *Learning Analytics*. Springer New York. ISBN 978-1-4614-3304-0 978-1-4614-3305-7; 2014, p. 61–75.
3. Lin, L.C., Prez, .A.J.. Educational Data Mining and Learning Analytics: differences, similarities, and time evolution. *RUSC Universities and Knowledge Society Journal* 2015;**12**(3):98–112.
4. Baker, R., Siemens, G.. Educational Data Mining and Learning Analytics. In: *The Cambridge Handbook of the Learning Sciences*; Cambridge Handbooks in Psychology. Cambridge University Press; 2014, .
5. Ferguson, R.. The State of Learning Analytics in 2012: A Review and Future Challenges. Tech. Rep.; Knowledge Media Institute; 2012.
6. Bhardwaj, B.K., Pal, S.. Data mining: A prediction for performance improvement using classification. *International Journal of Computer Science and Information Security* 2011;**9**(4):136–140.
7. Carnegie, D., Watterson, C., Andreae, P., Browne, W.. Prediction of success in engineering study. In: *2012 IEEE Global Engineering Education Conference (EDUCON)*. 2012, p. 1–9.
8. de Morais, A., Araujo, J., Costa, E.. Monitoring student performance using data clustering and predictive modelling. In: *2014 IEEE Frontiers in Education Conference (FIE)*. 2014, p. 1–8.
9. Guarin, L., Ernesto, C., Guzman, E.L., Gonzalez, F.A.. A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining. *Tecnologias del Aprendizaje, IEEE Revista Iberoamericana de* 2015;**10**(3):119–125.
10. Hashim, H., Talab, A.A., Satty, A., Talab, S.A.. Data mining methodologies to study students academic performance using the c4.5 algorithm. *International Journal of Computer Science and Information Security* 2015;**13**(4):104–113.
11. Al-Saleem, M., Al-Kathiry, N., Al-Osimi, S., Badr, G.. Mining Educational Data to Predict Students Academic Performance. In: Perner, P., editor. *Machine Learning and Data Mining in Pattern Recognition*; no. 9166 in Lecture Notes in Computer Science. Springer International Publishing; 2015, p. 403–414.
12. Quinlan, J.R.. Induction of decision trees. *Machine Learning* 1986;**1**(1):81–106.
13. Rokach, L., Maimon, O.. *Data Mining with Decision Trees: Theory and Applications*. World Scientific; 2014. ISBN 978-981-4590-09-9.
14. Home - RapidMiner Documentation. 2015 (accessed 2015-12-18). "http://docs.rapidminer.com/".